EA2022-descriptive-V3.R

Edre MA

2022-03-15

```
# ========================
# Descriptive Statistics
# R Biostat Workshop IIUM
# Edre MA, DrPH
# ========================


#You you are a researcher involved in a hypertension study
#objective 1: To describe the background characteristics of respondents
#objective 2: To determine the prevalence of hypertension
#objective 3: To determine the factors contributing to hypertension
```

```
#libraries needed to be installed
```

```
#readr - read csv file
#smartEDA - custom descriptive stat
#moments - skewness and kurtosis(normality)
#ggpubr - visualization of density (normal curve)
#usingR - histogram (normality)
#car - qqplot (normality)
#ggplot2 - visualization of boxplot
#dplyr - transform / mutate variables
#table1 - basic descriptive table

install.packages ("name of the package")
```

```
## Installing package into 'C:/Users/raef/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)

## Warning in install.packages :
##   package 'name of the package' is not available for this version of R
##
## A version of this package for your version of R might be available elsewhere,
## see the ideas at
## https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages
```

```
# data
```

```
#pulling the data from GitHub
```

```r
#go to https://github.com/adilzainal/IIUM_Biostatistic_workshop
#click "code" -> "Download ZIP"
#extract the ZIP file using WinRAR
#Create a new specific folder to store all files in your desktop
#set as working directory

#loading the data

#if csv (.csv)
install.packages("readr")
```

```
## Installing package into 'C:/Users/raef/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)

## package 'readr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpeQAkm8\downloaded_packages
```

```r
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```r
hstat <- read_csv("healthstatus6.csv") #load the file and make as object
```

```
## Rows: 153 Columns: 17

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): sex, exercise, smoking
## dbl (14): id, age, wt, ht, sbp, dbp, hba1c, hcy, wt2, wt3, sbp2, sbp3, dbp2...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
View(hstat)
```

```r
#objective 1: To describe the background characteristics of respondents
#summarising numerical values
```

```r
# we choose 3 IVs: age,sbp,dbp

install.packages("SmartEDA")
```

```
## Installing package into 'C:/Users/raef/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)

## package 'SmartEDA' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpeQAkm8\downloaded_packages
```

```
library(SmartEDA)
```

```
## Warning: package 'SmartEDA' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ExpCustomStat(hstat,
              Nvar=c("age","sbp","dbp"),
              stat = c('mean', 'sd', 'median', 'IQR'))
```

```
##    Attribute      mean       sd median IQR
## 1:       age  42.16340 8.932096     42  11
## 2:       sbp 132.24837 7.956527    132  13
## 3:       dbp  86.53595 6.268159     87   9
```

```
ExpCustomStat(hstat,
              Cvar=c("exercise","sex","smoking"),
              stat = c('count', 'prop'),gpby= FALSE)
```

```
##       Level Group_by count  prop
## 1: Moderate exercise    61 39.87
## 2:      Low exercise    74 48.37
## 3:     High exercise    18 11.76
## 4:     Male      sex    83 54.25
## 5:   Female      sex    70 45.75
## 6:      Yes  smoking    63 41.18
## 7:       No  smoking    90 58.82
```

```
#normality assumption check
```

```
#there are 5 criteria before you make decision what to report:

#1.mean~median

ExpCustomStat(hstat,
              Nvar=c("age","sbp","dbp"),
              stat = c('mean','median'))
```

```
##    Attribute      mean median
## 1:       age  42.16340     42
## 2:       sbp 132.24837    132
## 3:       dbp  86.53595     87
```

```
#2. acceptable skewness & kurtosis +-2d

install.packages("moments")
```

```
## Installing package into 'C:/Users/raef/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'moments' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpeQAkm8\downloaded_packages
```

```
library(moments)
ExpCustomStat(hstat,
              Nvar=c("age","sbp","dbp"),
              stat = c('skewness','kurtosis'))
```

```
##    Attribute    skewness kurtosis
## 1:       age  0.16179220 2.783220
## 2:       sbp  0.22172135 2.417301
## 3:       dbp -0.02148621 2.548945
```

```
#3. bell shaped curve (The MOST powerful determinant of normality)
install.packages("ggpubr")
```

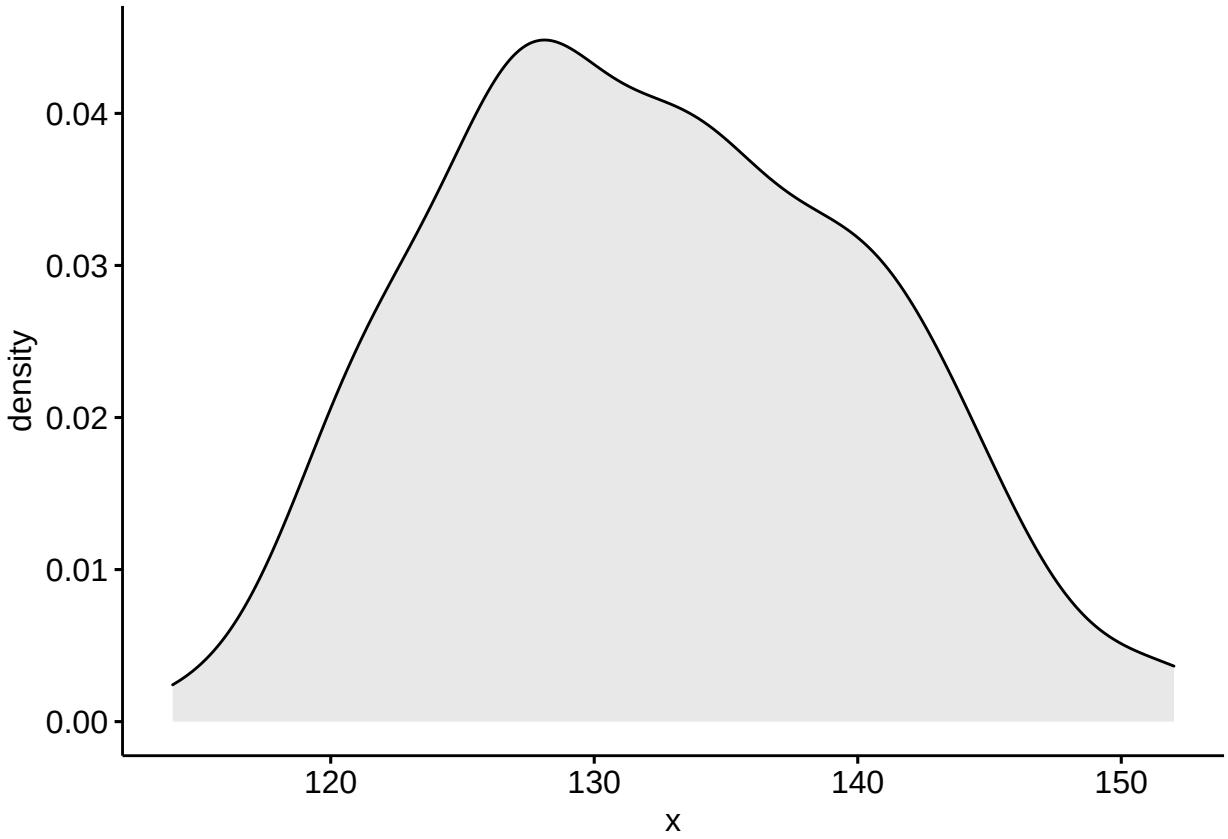
```
## Installing package into 'C:/Users/raef/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'ggpubr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpeQAkm8\downloaded_packages
```

```
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
ggdensity(hstat$sbp, fill = "lightgray")
```




```
install.packages("usingR")
```

```
## Installing package into 'C:/Users/raef/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
library(UsingR)
```

```
## Loading required package: MASS
```

```
## Loading required package: HistData
```

```
## Loading required package: Hmisc
```

```
## Loading required package: lattice
```

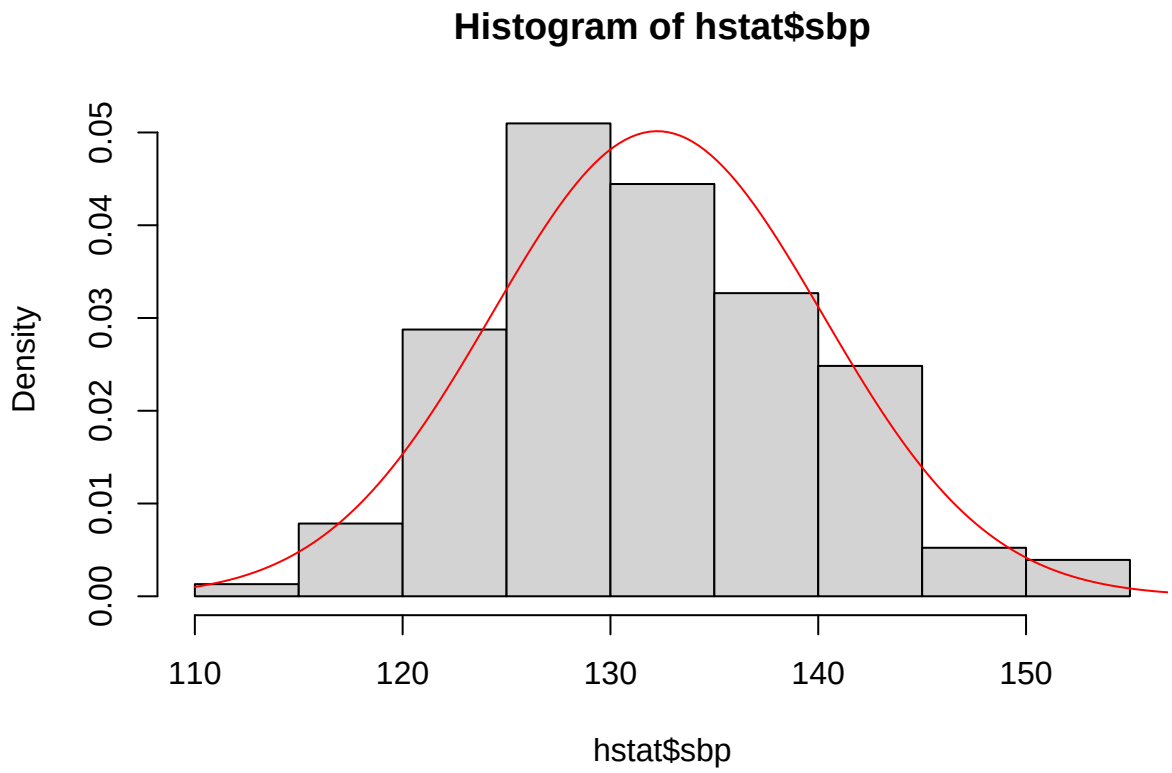
```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units

##
## Attaching package: 'UsingR'

## The following object is masked from 'package:survival':
##
##     cancer
```

```
hist(hstat$sbp, freq = FALSE)
x <- seq(110, 160, length.out=170)
y <- with(hstat, dnorm(x, mean(sbp), sd(sbp)))
lines(x, y, col = "red")
```



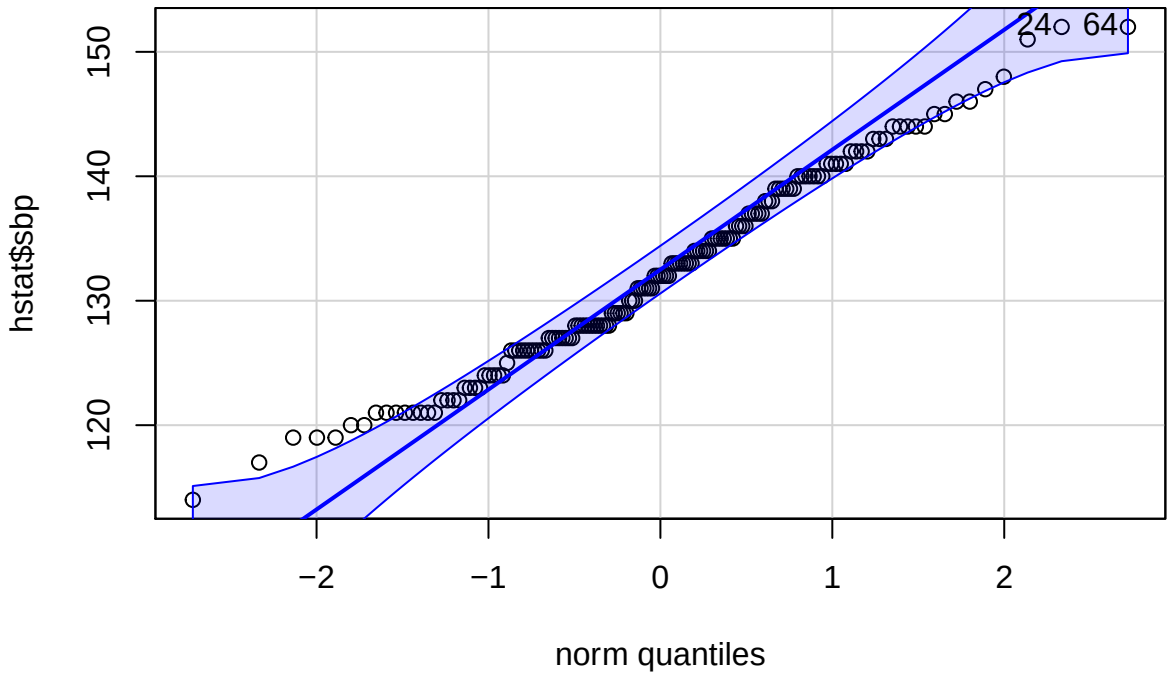

```
#4. qqplot

install.packages("car")
```

```
## Error in install.packages : Updating loaded packages
```

```
library(car)
```

```
## Loading required package: carData
```

```
qqPlot(hstat$sbp)
```




```
## [1] 24 64
```

```
#5. normality test
shapiro.test(hstat$sbp) #sample size less than 50
```

```
##
##  Shapiro-Wilk normality test
##
## data:  hstat$sbp
## W = 0.98403, p-value = 0.07418
```

```
ks.test(x, "pnorm", mean=mean(hstat$sbp), sd=sd(hstat$sbp))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.25558, p-value = 4.523e-10
## alternative hypothesis: two-sided
```

```
#finally, make your decision
ExpCustomStat(hstat,
              Nvar=c("age","sbp","dbp"),
              stat = c('mean', 'sd'))
```

```
##    Attribute      mean       sd
## 1:       age  42.16340 8.932096
## 2:       sbp 132.24837 7.956527
## 3:       dbp  86.53595 6.268159
```

```
#summarising categorical values
```

```
ExpCustomStat(hstat,
              Cvar=c("sex", "smoking"),
              gpby=FALSE)
```

```
##     Level Group_by Count  Prop
## 1:   Male      sex    83 54.25
## 2: Female      sex    70 45.75
## 3:    Yes  smoking    63 41.18
## 4:     No  smoking    90 58.82
```

```
#count refers to the frequency, n
#proportion here refers to the percentage distribution of that category
```

```
#missing data
```

```
#usually coded as "NA" in the dataset
#we create a dummy object first to showcase this exercise
missing <- hstat
missing[missing$id==57, "sbp"] <- NA

#demonstrating the row to show the missing value using dummy data
missing$sbp
```

```
##   [1] 123 122 136 127 151 128 146 145 134 122 124 138 127 145 138 126 122 128
##  [19] 135 117 147 135 139 152 126 121 132 139 137 144 135 141 130 131 144 129
##  [37] 126 127 136 123 124 121 127 131 134 124 139 128 127 132 143 128 130 144
##  [55] 124 141  NA 135 121 140 142 128 146 152 144 142 132 137 126 133 128 141
##  [73] 126 119 125 130 131 140 123 120 127 126 119 140 121 134 133 131 129 128
##  [91] 140 139 143 129 126 133 136 128 134 132 140 137 140 135 127 128 128 143
## [109] 133 119 126 132 133 131 126 140 136 135 128 141 139 135 137 132 114 121
## [127] 122 121 142 133 133 142 129 129 141 129 139 148 121 133 131 128 144 134
## [145] 123 126 120 138 135 127 124 134 121
```

```
which (is.na(missing$sbp))
```

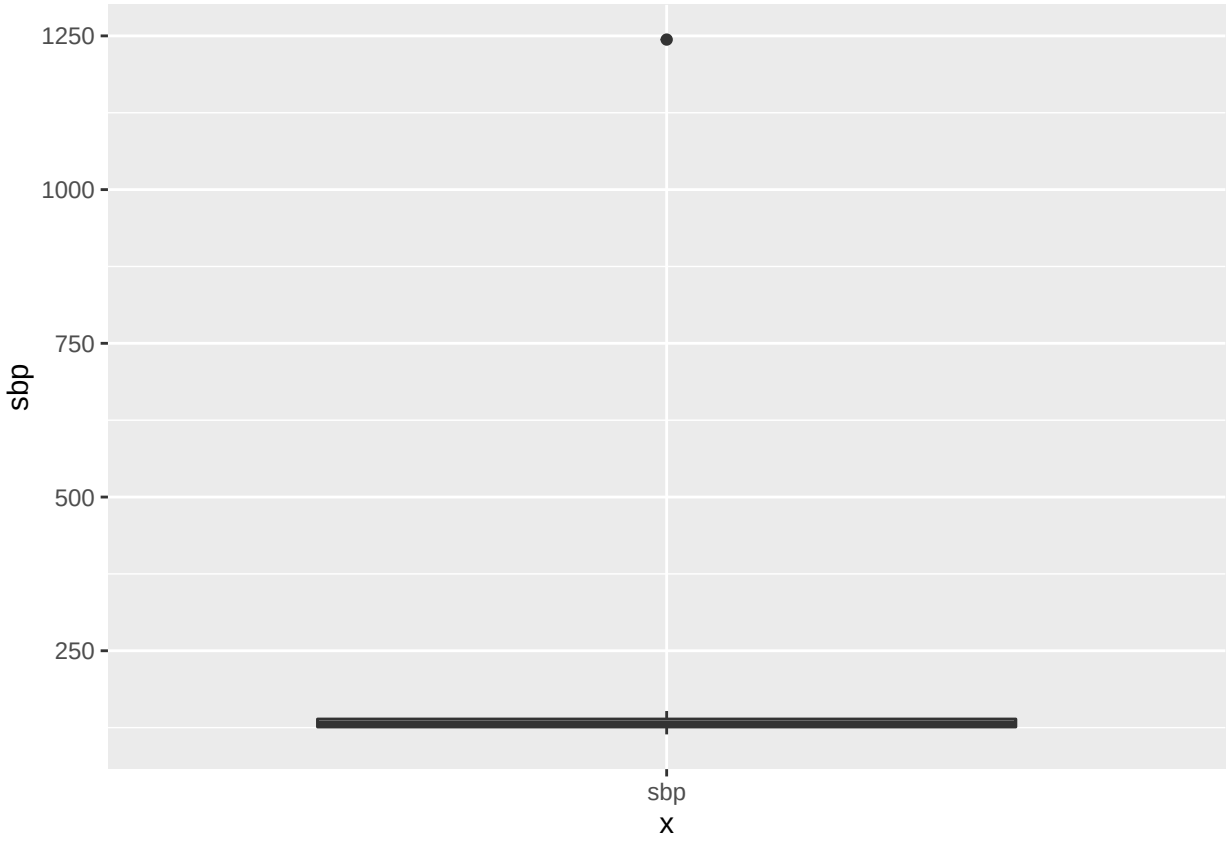
```
## [1] 57
```

```r
#outlier detection
```

```r
#create an outlier dummy data
outlierdummy <- hstat
outlierdummy[outlierdummy$id==131, "sbp"] <- 1244

#visual method
install.packages("ggplot2")
```

```
## Error in install.packages : Updating loaded packages
```

```r
library(ggplot2)
ggplot(outlierdummy, aes(x = "sbp", y = sbp)) + geom_boxplot()
```




```r
#data row method

is_outlier <- outlierdummy$sbp > 250 | outlierdummy$sbp < 70
is_outlier
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
#objective 2: To determine the prevalence of hypertension
#objective 3: To determine the factors contributing to hypertension
#basic data transformation:categorizing
```

```
install.packages("dplyr")
```

```
## Error in install.packages : Updating loaded packages
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:Hmisc':
##
##     src, summarize

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#hypertension status (either sbp or dbp equal or more than 140/90mmHg, respectively, considered hyperte
#to answer objective 2

hstat2 <-hstat %>%
  mutate(hpt=if_else(hstat$sbp<140 & hstat$dbp<90,'normal','high'))

View(hstat2)

ExpCustomStat(hstat2,
              Cvar="hpt",
              stat=c("count","prop"))
```

```
##       hpt count  prop
## 1: normal    94 61.44
## 2:   high    59 38.56
```

```
#to make data preparation for objective 3
#glucose control (6.5% and above considered poor)


hstat2$glucontrol<-cut(hstat2$hba1c,
                           breaks=c(-Inf,6.49,Inf),
                           labels=c("Good", "Poor"))
summary(hstat2)
```

```
##        id            age            sex              exercise
##  Min.   :  1   Min.   :21.00   Length:153         Length:153
##  1st Qu.: 39   1st Qu.:36.00   Class :character   Class :character
##  Median : 77   Median :42.00   Mode  :character   Mode  :character
##  Mean   : 77   Mean   :42.16
##  3rd Qu.:115   3rd Qu.:47.00
##  Max.   :153   Max.   :64.00
##    smoking                wt              ht             sbp
##  Length:153         Min.   :42.60   Min.   :140.0   Min.   :114.0
##  Class :character   1st Qu.:55.40   1st Qu.:148.0   1st Qu.:126.0
##  Mode  :character   Median :59.10   Median :156.0   Median :132.0
##                     Mean   :60.92   Mean   :155.8   Mean   :132.2
##                     3rd Qu.:64.20   3rd Qu.:162.0   3rd Qu.:139.0
##                     Max.   :82.00   Max.   :176.0   Max.   :152.0
##       dbp             hba1c            hcy             wt2
##  Min.   : 71.00   Min.   : 2.400   Min.   : 8.80   Min.   :39.59
##  1st Qu.: 82.00   1st Qu.: 5.800   1st Qu.:12.60   1st Qu.:52.09
##  Median : 87.00   Median : 7.100   Median :14.20   Median :55.76
##  Mean   : 86.54   Mean   : 7.048   Mean   :15.08   Mean   :58.19
##  3rd Qu.: 91.00   3rd Qu.: 8.300   3rd Qu.:16.10   3rd Qu.:62.57
##  Max.   :100.00   Max.   :11.000   Max.   :42.00   Max.   :81.54
##       wt3              sbp2            sbp3            dbp2
##  Min.   :39.43   Min.   :113.0   Min.   :111.0   Min.   : 62.00
##  1st Qu.:51.25   1st Qu.:125.0   1st Qu.:125.0   1st Qu.: 77.00
##  Median :55.11   Median :131.0   Median :130.0   Median : 82.00
##  Mean   :57.61   Mean   :131.6   Mean   :130.7   Mean   : 82.31
##  3rd Qu.:61.85   3rd Qu.:138.0   3rd Qu.:137.0   3rd Qu.: 87.00
##  Max.   :81.07   Max.   :152.0   Max.   :153.0   Max.   :102.00
##       dbp3            hpt            glucontrol
##  Min.   :67.00   Length:153         Good: 51
##  1st Qu.:76.00   Class :character   Poor:102
##  Median :81.00   Mode  :character
##  Mean   :81.15
##  3rd Qu.:86.00
##  Max.   :98.00
```

```
#bmistatus (WHO classification)

hstat3<- hstat2 %>%
  mutate(height_m = ht / 100,bmi = wt / (height_m^2))
```

```
View(hstat3)

hstat3$bmistatus<- cut(hstat3$bmi,
                           breaks=c(-Inf, 18.49999, 24.9999, 29.9999, Inf),
                           labels=c("underweight", "normal", "overweight", "obese"))
summary(hstat3)
```

```
##        id            age            sex              exercise
##  Min.   :  1   Min.   :21.00   Length:153         Length:153
##  1st Qu.: 39   1st Qu.:36.00   Class :character   Class :character
##  Median : 77   Median :42.00   Mode  :character   Mode  :character
##  Mean   : 77   Mean   :42.16
##  3rd Qu.:115   3rd Qu.:47.00
##  Max.   :153   Max.   :64.00
##    smoking                wt              ht             sbp
##  Length:153         Min.   :42.60   Min.   :140.0   Min.   :114.0
##  Class :character   1st Qu.:55.40   1st Qu.:148.0   1st Qu.:126.0
##  Mode  :character   Median :59.10   Median :156.0   Median :132.0
##                     Mean   :60.92   Mean   :155.8   Mean   :132.2
##                     3rd Qu.:64.20   3rd Qu.:162.0   3rd Qu.:139.0
##                     Max.   :82.00   Max.   :176.0   Max.   :152.0
##       dbp             hba1c            hcy             wt2
##  Min.   : 71.00   Min.   : 2.400   Min.   : 8.80   Min.   :39.59
##  1st Qu.: 82.00   1st Qu.: 5.800   1st Qu.:12.60   1st Qu.:52.09
##  Median : 87.00   Median : 7.100   Median :14.20   Median :55.76
##  Mean   : 86.54   Mean   : 7.048   Mean   :15.08   Mean   :58.19
##  3rd Qu.: 91.00   3rd Qu.: 8.300   3rd Qu.:16.10   3rd Qu.:62.57
##  Max.   :100.00   Max.   :11.000   Max.   :42.00   Max.   :81.54
##       wt3              sbp2            sbp3            dbp2
##  Min.   :39.43   Min.   :113.0   Min.   :111.0   Min.   : 62.00
##  1st Qu.:51.25   1st Qu.:125.0   1st Qu.:125.0   1st Qu.: 77.00
##  Median :55.11   Median :131.0   Median :130.0   Median : 82.00
##  Mean   :57.61   Mean   :131.6   Mean   :130.7   Mean   : 82.31
##  3rd Qu.:61.85   3rd Qu.:138.0   3rd Qu.:137.0   3rd Qu.: 87.00
##  Max.   :81.07   Max.   :152.0   Max.   :153.0   Max.   :102.00
##       dbp3            hpt            glucontrol    height_m          bmi
##  Min.   :67.00   Length:153         Good: 51   Min.   :1.400   Min.   :15.65
##  1st Qu.:76.00   Class :character   Poor:102   1st Qu.:1.480   1st Qu.:22.06
##  Median :81.00   Mode  :character              Median :1.560   Median :24.89
##  Mean   :81.15                                 Mean   :1.558   Mean   :25.31
##  3rd Qu.:86.00                                 3rd Qu.:1.620   3rd Qu.:28.22
##  Max.   :98.00                                 Max.   :1.760   Max.   :38.88
##        bmistatus
##  underweight: 6
##  normal     :75
##  overweight :48
##  obese      :24
##
##
```

```
#Reporting your descriptive analysis
```

```
install.packages("table1")
```

```
## Installing package into 'C:/Users/raef/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'table1' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpeQAkm8\downloaded_packages
```

```
library(table1)
```

```
##
## Attaching package: 'table1'
```

```
## The following objects are masked from 'package:Hmisc':
##
##     label, label<-, units
```

```
## The following objects are masked from 'package:base':
##
##     units, units<-
```

```
table1(~  age + factor(smoking) + factor(exercise) + wt + bmi, data=hstat3)
```

```
## Get nicer `table1` LaTeX output by simply installing the `kableExtra` package
```

	Overall
	(N=153)
age	
Mean (SD)	42.2 (8.93)
Median [Min, Max]	42.0 [21.0, 64.0]
factor(smoking)	
No	90 (58.8%)
Yes	63 (41.2%)
factor(exercise)	
High	18 (11.8%)
Low	74 (48.4%)
Moderate	61 (39.9%)
wt	
Mean (SD)	60.9 (8.27)
Median [Min, Max]	59.1 [42.6, 82.0]
bmi	
Mean (SD)	25.3 (4.29)
Median [Min, Max]	24.9 [15.6, 38.9]

```
#thank you
```